MANCHESTER 1824

The University of Manchester

M\cr NA
Manchester Numerical Analysis

Statistinis Apvalinimas ir Nauji Skaičių Formatai Superkompiuteriuose

Mantas Mikaitis
Matematikos Departamentas
Mančesterio Universitetas

`mantas.mikaitis@manchester.ac.uk`

10-asis Lietuvos Jaunųjų Matematikų Susitikimas

Gruodžio 28 d. 2021, Matematikos ir Informatikos Fakultetas, Vilniaus Universitetas, Vilnius, Lietuva

Kas yra Kompiuterių Aritmetika?

- **Skaičių vaizdavimas kompiuteryje**: turime registrus kurie laiko bitus (0/1)—kaip pavaizduoti realius skaičius?
- **Operacijos su skaičiais**: $+, -, \times, \div, \sqrt{}$ ir kt.
- **Specialios funkcijos**: $e^x, \log_e x, \sin, \cos$ ir kt.
- **Apvalinimas**: link artimiausio skaičiaus, link nulio, link begalybes ir kt.

Bendrai

Dalies **realių skaičių** vaizdavimas ir skaičiavimas su jais.

Kompiuterių Aritmetikos Tyrimai

Tyrimai šitoje srityje gali buti: **programinė ir techninė įranga**, **bendri algoritmai aritmetikai**, **matematinė analizė**, paklaidų nagrinėjimas programose ir kt.

Slankiojo kablelio aritmetika

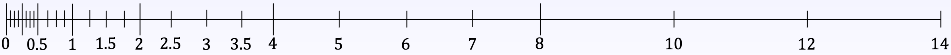
Ribota sistema $F = F(\beta, t, e_{min}, e_{max}) \in \mathbb{R}$ kurioje kiekvienas elementas turi išraišką

$$x = \pm m \times \beta^{e-p+1}.$$

Kompiuteriuose dažniausiai sutinkam $\beta = 2$.
p nusako **formato tikslumą**, $e_{min} \leq e \leq e_{max}$ yra **skaičiaus eilė**, o $m \leq \beta^p - 1$ yra **mantise** (p, e, ir m sveikieji skaičiai).

Pavyzdys

Apačioje—teigiami skaičiai sistemoje
$F(\beta = 2, p = 3, e_{min} = -2, e_{max} = 3)$.



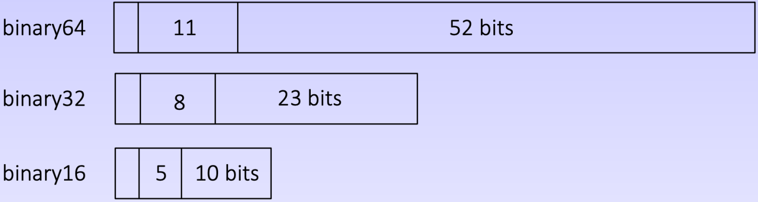
IEEE 754 Standartinė Aritmetika

- Standartas skirtas **suvienodinti aritmetikas tarp kompiuterių** (IEEE (2019)).
- Išleistas 1985, atnaujintas 2008 and 2019.
- Rekomenduojami **formatai**, **operacijos**, **apvalinimo metodai** ir kt.
- Dauguma šiodieninių kompiuterių palaiko.

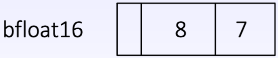
Table: Formatai ($\beta = 2$) iš IEEE 754 standardo. $f_{\min}$—mažiausias normalizuotas skaičius, $s_{\min}$—mažiausias nenormalizuotas skaičius, $f_{\max}$—didžiausias skaičius.

	binary16	binary32	binary64
p	11	24	53
$f_{\min}$	2^{-14}	2^{-126}	2^{-1022}
$s_{\min}$	2^{-24}	2^{-149}	2^{-1074}
$f_{\max}$	$2^{15}(2-2^{-10})$	$2^{127}(2-2^{-23})$	$2^{1023}(2-2^{-52})$

Kompiuterių atmintyje IEEE formatai laikomi naudojant bitus (**binarinė sistema**):



Taippat naujas **nestandatinis bfloat16 formatas**, naudojamas neuroniniuose tinkluose:



TOP500 Superkompiuteriai ir jų Aritmetika

- **TOP500** (`https://www.top500.org/`) superkompiuterių (HPC) sarašas atnaujinamas kas pusę metų.
- **143 kompiuteriai turi NVIDIA vaizdo plokštes** (nuo 2016 našumo dalis išaugo nuo 12% iki 39%).
- Prieš dešimtmetį **binary16** nebuvo galima rasti, dabar vis daugėja.

Figure: **NVIDIA** vaizdo plokštė.

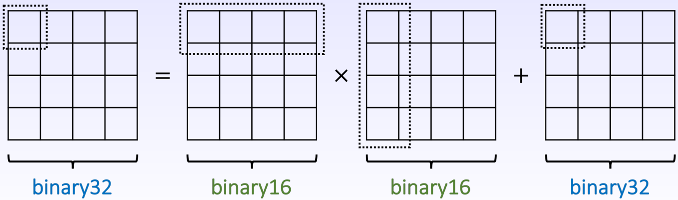
Figure: Top 2 **Summit** kompiuteris.

Figure: Top 1 **Fugaku** kompiuterio blokas.

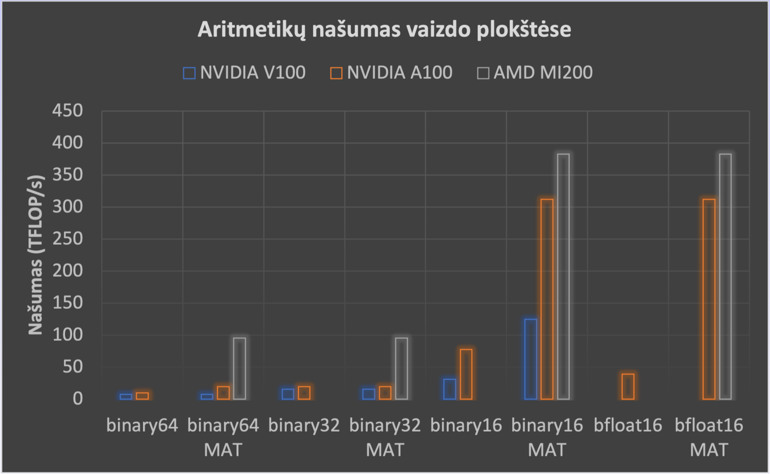
TOP500 Superkompiuteriai ir jų Aritmetika

- Taippat **naujo tipo operacijos su matricomis**: "**tensor core**" (TC).
- Priima argumentus **binary16** matricas ir jas padaugina.
- Gražina matricas **binary32**.
- Žymiai greičiau negu dauginti atskirai po elementą.

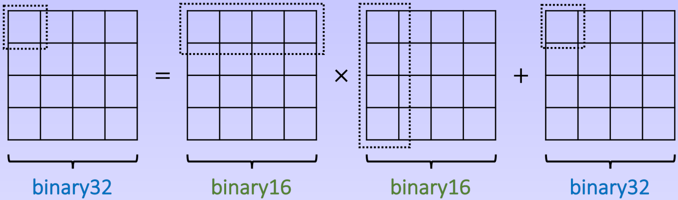
Turint binary16 matricas $A \in R^{4\times4}$ ir $B \in R^{4\times4}$, binary32 matricą $C \in R^{4\times4}$, TC skaičiuoja $D = AB + C$ (**64 sudėties-daugybos operacijos vienu metu**).



TOP500 Superkompiuteriai ir jų Aritmetika



Aritmetikos Testavimas



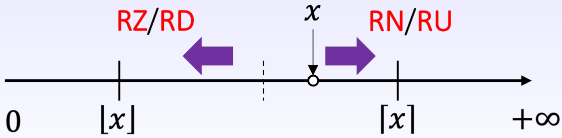
$$\overbrace{d_{11}}^{\text{binary32}} = \underbrace{\underbrace{\underbrace{a_{11}}_{\text{binary16}} b_{11}}_{\text{Exact mult. (not rounded to binary16): 22 signif. bits, 6 exp. bits, 1 sign bit}} + a_{12}b_{21} + a_{13}b_{31} + a_{14}b_{41} + \underbrace{c_{11}}_{\text{binary32}}}_{\text{binary32 5-operand adder}}$$

Aritmetikos Testavimas

Pagrindinis klausimas: ar šitie prietaisai matricoms dauginti užtikrintai veikia taippat kaip programinė įranga (**pagal IEEE 754**)?

Pavyzdžiui, galim patikrinti:

- Kokia seka **keturios sudėties operacijos** yra atliekamos?
- Kaip apvalinama sudėtyje?
- Daugiau: Fasi, Higham, Mikaitis, Pranesh (2021)



Aritmetikos Testavimas

Pasirenkam skaitinę savybę testavimui

Pavyzdžiui: Apvalinimas

⇩

Pasirenkam kokie galimi variantai

Kandidatai: RU/RD

⇩

Pasirenkam skaičius kurie padauginus duotų skirtingus atsakymus kiekvienam apvalinimo metodui

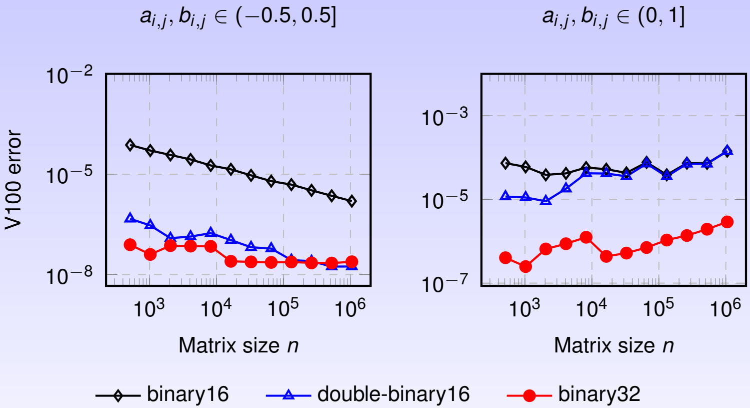
Pasirenkam skaičių kuris netelpa į slankiojo kablelio formatą

⇩

Leidžiam skaičius per sistemą ir tikrinam ką gaunam pagal kandidatus

Atsakymas kairėje: RD; dešnėje: RU.

Paklaidos Matricų Daugyboje



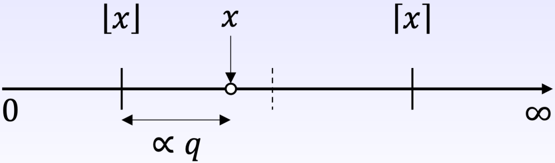
Rezultatai: Fasi, Higham, Lopez, Mary, Mikaitis (2022).

Statistinis Apvalinimas

Apibrėžimas iš Connolly, Higham, Mary (2021).
Jeigu $x \in \mathbb{R}$, $\lfloor x \rfloor \leq x \leq \lceil x \rceil$ (x tarp dvieju slankiojo kablelio skaičių), **statistinis apvalinimas** apibrėžiamas kaip

$$\mathrm{SR}(x) = \begin{cases} \lceil x \rceil \textbf{su tikimybe } q(x), \\ \lfloor x \rfloor \textbf{su tikimybe } 1 - q(x), \end{cases} \tag{1}$$

ir $q(x) = \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}$.



Statistinis Apvalinimas

- SR **nėra dažnai sutinkamas kompiuteriuose** (nors yra tarp Intel ir Graphcore tam tikrų mikroprocesorių).
- Minėtas publikacijose iš 1950, bet tik neseniai vėl susidomėta neuroninių tinklų literatūroje.
- **Statistiškai naudingas apvalinimo metodas**: neapvalina į vieną pusę net jeigu visi skaičiai arčiau kairiojo slankiojo kablelio skaičiaus.

Tyrimai ties Statistiniu Apvalinimu

Aritmetikos su SR teorija, **paklaidos** mokslinėse simuliacijose, **SR algoritmai** programinei ir techninei įrangai su SR mikroprocesoriouse.

Euler Metodas su Statistiniu Apvalinimu

Spredžiam dvi diferencialines lygtis su Euler metodu:

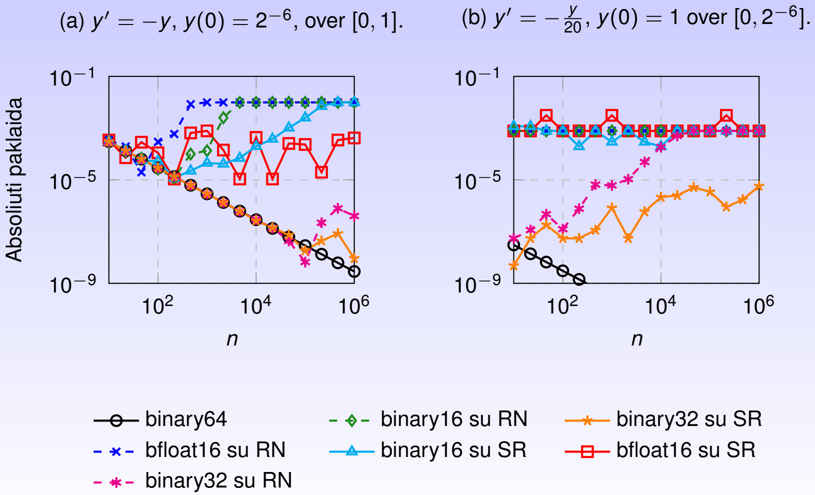
- $y_{n+1} = y_n - hy_n$, su $y_0 = 2^{-6}$, intervale $[0, 1]$ su žingsniais $h = 1/n$.
- $y_{n+1} = y_n - h\frac{y_n}{20}$, su $y_0 = 1$, intervale $[0, 2^{-6}]$ su žingsniais $h = 2^{-6}/n$.

Eksperimentas keičiant n

Didinsim $n \in [10, 10^6]$ tol kol metodo žingsnis pasidarys mažesnis už tarpus tarp skaičių slankiojo kablelio aritmetikoje, kur ir pasimatys apvalinimo aritmetikoje poveikis paklaidoms.

Daugiau info: Croci, Fasi, Higham, Mary, Mikaitis (2021), Fasi, Mikaitis (2021).

Rezultatai: Fasi, Mikaitis (2021).



Euler Metodas su Statistiniu Apvalinimu

Kitas pavyzdys. Spredžiam

$$u'(t) = v(t),\ v'(t) = -u(t)$$

kai $u(0) = 1,\ v(0) = 0$ (**vienetinis skritulys** uv koordinačių plokštėje).

Sprendimas naudojant Euler metodą (žingsniai $h = 2\pi/n$):

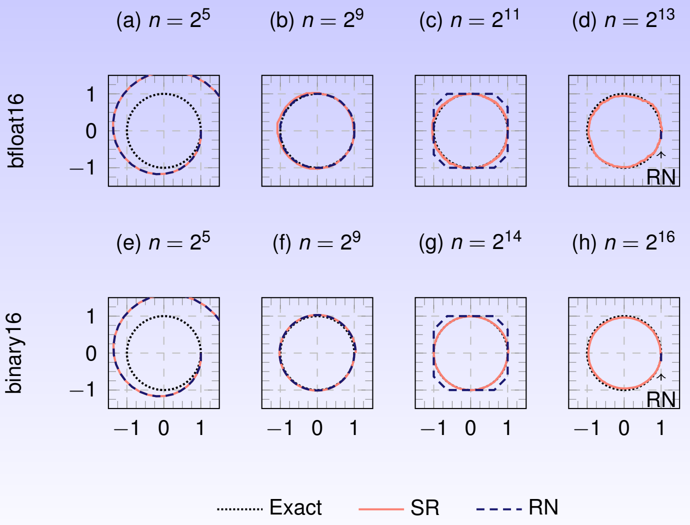
$$u_{k+1} = u_k + hv_k,\ v_{k+1} = v_k - hu_k.$$

Eksperimentas mažinant h

Didinam n kol apvalinimo paklaidos dominuoja.

Daugiau info: Croci, Fasi, Higham, Mary, Mikaitis (2021), Fasi, Mikaitis (2021).

Euler Metodas su Statistiniu Apvalinimu



Įvairių Aritmetikų Simuliacija ant Kasdieninių Kompiuterių

- Jeigu neturim priegos prie naujų aritmetikų, ka darom?
- **Simuliuojam aritmetikas** ant stalinių kompiuterių.
- Ant `MATLAB` galim naudoti funkciją `chop` (Higham, Pranesh (2019)).
- Tarp `C` kalbos yra `CPFloat` funkcija (Fasi, Mikaitis (2020))—veikia ir tarp `MATLAB`.
- Palaiko betkokią aritmetiką $p \leq 32$.
- Palaiko visus apvalinimo metodus, įskaitant statistinį.
- **Naudoja binary64 ant procesoriaus operacijoms** ir tada nunuliuoja kiek reikia bitų dešnėje.

binary64		11	52 bits

MATLAB CPFloat pavyzdys

```
>> options.format = 'binary16';
>> x = pi;
>> x
x = 3.14159265358979

>> xr = cpfloat(x, options);
>> xr
xr = 3.140625

>> x*x
ans = 9.86960440108936

>> cpfloat(x*x, options)
ans = 9.8671875
```

Santrauka

- Vis dažniau atsiranda **binary16**, **bfloat16** ir matricų operacijos kompiuterių mikroprocesoriuose.
- Programinė įranga adaptuojama **dėl geresnio našumo**.
- Pristačiau progresą testuojant ir simuliuojant įvairias aritmetikas.
- Skaidrės `https://mmikaitis.github.io/talks/`.

Visada ieškau su kuo bendradarbiauti tyrimuose šioje srityje. Susiekimui mantas.mikaitis@manchester.ac.uk.

References I

IEEE
IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2019 (revision of IEEE Std 754-2008).
Institute of Electrical and Electronics Engineers, Piscataway, NJ, USA. 2019.

M. Fasi, N. J. Higham, M. Mikaitis, and S. Pranesh.
Numerical Behavior of NVIDIA Tensor Cores.
PeerJ Comput. Sci. 7:e330 (2021).

M. Fasi, N. J. Higham, F. Lopez, T. Mary, and M. Mikaitis.
Multiword Matrix Multiplication: General Error Analysis and Application to GPU Tensor Cores.
2022. In preparation.

References II

M. P. Connolly, N. J. Higham, and T. Mary.
Stochastic Rounding and Its Probabilistic Backward Error Analysis.
SIAM J. Sci. Comput. 43, A566–A585. 2021.

M. Croci, M. Fasi, N. J. Higham, T. Mary, and M. Mikaitis.
Stochastic Rounding: Implementation, Error Analysis, and Applications.
MIMS EPrint 2021.17. 2021.

N. J. Higham and S. Pranesh.
Simulating Low Precision Floating-Point Arithmetic.
SIAM J. Sci. Comput. 41, C585–C602. 2019.

References III

M. Fasi and M. Mikaitis.
CPFloat: A C Library for Emulating Low-Precision Arithmetic.
MIMS EPrint 2020.22. 2020.

M. Fasi and M. Mikaitis.
Algorithms for Stochastically Rounded Elementary Arithmetic Operations in IEEE 754 Floating-Point Arithmetic.
IEEE Trans. Emerg. Topics Comput. 9, 1451–1466. 2021.